REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE

3. REPORT TYPE AND DATES COVERED
FINAL 01 SEP 93 TO 31 AUG 96

4. TITLE AND SUBTITLE
AASERT93- WEATHER ANALYSIS AND PREDICTION USING EMPIRICAL ORTHOGONAL FUNCTIONS

5. FUNDING NUMBERS
F49620-93-1-0531

6. AUTHOR(S)
RICHARD L. PFEFFER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
GEOPHYSICAL FLUID DYNAMICS INSTITUTE
FLORIDA STATE UNIVERSITY
TALLAHASSEE, FL 32306-3017

8. PERFORMING ORGANIZATION REPORT NUMBER
AFOSR-TR-96
0482

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
110 DUNCAN AVE, SUITE B115
BOLLING AFB DC 20332-8080

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE:
DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
SEE REPORT

19961016 126

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

FINAL REPORT

Under

Air Force Office of Scientific Research
AASERT Grant F49620-93-1-0531

Statistical Weather Prediction

by

Richard L. Pfeffer
Principal Investigator
Geophysical Fluid Dynamics Institute
Florida State University
Tallahassee FL 32306-3017

August 1996

1. Introduction

The primary purpose of AASERT Grant F49620-93-1-0531 was to support graduate student Scott Applequist's efforts to explore statistical methods of weather prediction under the supervision of Professor Pfeffer. The research Scott has done under this grant has enabled him to formulate his Ph. D. dissertation on cold season regional weather prediction in the range of 6 to 36 hours. His dissertation will be completed under AFOSR Grant F49620-96-1-0172.

Owing to the availability of plentiful surface and upper air data in both space and time over the eastern portion of the United States, Scott has chosen to test various *nonlinear* statistical methodologies over this region. Once successful methodologies are found, they can be tested over areas with more sparse data coverage. Scott's goal is to ascertain the extent to which nonlinear statistical methods can improve upon linear regression techniques and upon the accuracy of current weather forecasts.

2. Statistical Methodologies in Use Today

The two main statistical forecasting methods in use in meteorology today are both based on linear regression. One is the "perfect prog" technique and the other the "model output statistics" (MOS) technique (Wilks 1995). The first develops linear relationships between synoptic variables and local weather events based on a long record of observational data and then uses numerical forecasts of the synoptic variables as the predictors in these relationships, treating them as if they were actual data. The second uses linear regression to relate numerical model output directly to local weather events, correcting the model biases in the process. While the MOS technique is very appealing in principal, it is difficult to implement, because, in order to generate the linear regression coefficients, it requires the use of a numerical prediction model that remains fixed over a period of several years, longer than most models are kept in service because of ongoing improvements in numerical methods and parameterizations.

3. Testing Nonlinear Methods

Scott's approach is to specify the synoptic variables, both at the surface and aloft, at three-hourly intervals and then use them as predictors in statistical forecasts of the actual weather at an array of individual locations. His efforts under the AASERT grant began with a study of how best to specify the synoptic variables above the surface of the earth at three-hourly intervals.

On the basis of his analysis of a long record of observational data, Scott ascertained that the 850-mb moisture and temperature advections were important precursors to widespread precipitation events. But observational data at 850-mb are available only at 12-hour intervals, which is too far apart to be useful in short-range weather prediction. Since a 30-year record of surface data is available over the United States at three-hour intervals, and since surface and 850mb patterns are strongly related, he tried using the surface pressure as a predictor of 850-mb geopotential height. The plan was to test this idea first on predicting the 850-mb geopotential field and, if successful, to extend it to temperature, wind and water vapor, from which the advections would be calculated. He tested two methods for doing this. In one, he resolved the data at both levels into EOFs and their corresponding PCs and used the EOFs corresponding to the surface data as predictors of the three-hourly patterns at 850-mb. In the other he used canonical correlation analysis (CCA) (Wilks 1995) to relate the two fields. In some cases the spatial anomaly correlations between the predicted and verification fields were 0.9 or better for the dependent data set. But, in comparing these results with 12-hour predictions of the same fields by numerical models, he found that the numerical model forecasts showed greater accuracy and consistency than did the statistical techniques. Recognizing that such numerical forecasts would be readily available for use as predictors in statistical forecasting of weather at individual stations, he shifted the focus of his research to the latter task.

Scott chose to use the perfect prog method. The first task he faced was to find a method of selecting synoptic variables from the numerical model to use as predictors. Given a pool of 50 predictors (including time sequences of temperature, pressure, wind, specific humidity, potential vorticity, etc., as well as the nonlinear advections of these variables prior to the forecast time), it would not be feasible to test all possible combinations of these predictors to select the best 10. Such a process would require comparing 10^{10} combinations, which Scott estimated would take a computer 33 years to complete. Instead, he chose a much quicker (although approximate) way to screen the predictors that has been proven to give results that are close to the optimum combination. It is called "forward selection" or "stepwise regression" (Wilks 1995). In this methodology, one first loops once through the data using a linear regression model to find the single best predictor x_1 which minimizes the error of the predictand in the least squares sense. Then, from among the remaining predictors, one follows the same procedure to find the best predictor x_2 which minimizes the *remaining* error, and so on until the desired number of predictors has been reached. For the purpose of finding the best 10 out of 50 possible predictors, this procedure requires only 50+49+...+41 = 455 examinations, which is quite feasible to do. Glahn (1985) suggested selecting predictors in this manner until one of three criteria has been met, viz., (*i*) the reduction of error becomes less than 0.5% of the variance of the parameter to be predicted, (*ii*) the number of predictors reaches 12, or (*iii*) an F-test shows that an additional predictor does not significantly reduce the error further. While the skill in predicting the variables within a training set of data may continue to increase as predictors are added to the pool, Lorenz (1977) has shown that, when applied to an *independent* data set, the skill usually decreases after a certain number of predictors has been reached. Accordingly, in applying the forward selection methodology, Scott is testing his forecasts on independent data before deciding on the final number of predictors to use.

Once the predictors have been chosen, it is necessary to decide on the statistical method to use in the prediction. The focus of Scott's research is the testing of three methods for incorporating nonlinearity and comparing them against straightforward linear regression, which is being used as the baseline. The first is the inclusion of nonlinear advections at several time steps prior to the prediction time in the pool of predictors in a linear regression scheme. Although the scheme is linear, nonlinearity is incorporated in two ways, first in the use of nonlinear advections as predictors, and secondly (as demonstrated by Lorenz 1977) in the use of a time sequence of predictors.

The second method involves the use of "neural networks". A neural network is an artificial intellegence scheme (intended to model the behavior of neurons in the human brain) in which the predictors in a training set of data are used to exite changes in the numerical values assigned to different nodes. The number of intermediate (or "hidden") nodes separating the predictors from the predictands is determined by the user, and the weighting factors assigned to each node are computed empirically via the method of "back propagation" (Lawrence 1991). In this method the error signal is fed back through the network, altering the weights until the error has been reduced below a desired threshold. Nonlinearity is incorporated by the use of a sigmoidal transfer function (usually a hyperbolic tangent) that produces a nonlinear response to a linear input.

The third methodology is the "classifier system". This is an expert system in which the predictands are altered in successive steps according to prescribed rules that relate them to the predictors. Scott is using a "genetic" search algorithm (Goldberg, 1989) to determine the best rules that relate the predictors to predictands such as local precipitation and cloud cover. Genetic search algorithms mate generations of rules to one another until the optimum set emerges. These rules are characteristically in the form of a set of threshold values of the predictor beyond which a specified amount of the predictand (for example, precipitation) is added to the prediction.

4. A Trial Using the Lorenz Equations

Before attempting to employ the procedures on 30 years of surface and upper air data at a large array of observing stations over the eastern U.S., it seemed prudent to determine in advance what problems might arise in the implementation of the different methodologies, and whether the use of nonlinear statistical methods can be expected to make a significant difference in forecasts of chaotic behavior. Accordingly, Scott chose to test the methodologies he selected on the prediction of the three variables in the low order Lorenz (1963) equations for atmospheric convection. Starting with initial conditions for the three variables x, y and z that are nearly on the attractor, he integrated these equations to generate a sequence of 52,000 values of the variables, each separated from the preceding one by a nondimensional time increment of .01. This value is a very small fraction of an orbital period. He discarded the first 2,000 values in order to eliminate transient effects arising from the fact that the initial conditions do not lie exactly on the attractor.

Inspection of the autocorrelation function corresponding to each variable revealed monotonic decreases with time lag for x and y (with the latter decreasing faster) and a quasi-periodic fluctuation for z. At a time lag of .25 nondimensional units, the autocorrelations for x, y and z had decreased to 0.5, 0.35 and -0.30, respectively. These values were deemed sufficiently low that a prediction made for .25 time units into the future would not benefit much from persistence or linear extrapolation. Accordingly, statistical forecasts were prepared for periods of .25 time units and compared with the solution of the Lorenz equations. In order to speed up the calculations and test the stationarity of the prediction coefficients, the output was parcelled into blocks of 5,000 time steps, each block corresponding to roughly 60 orbits in the attractor. Within this time frame the trajectory typically switches between positive and negative values of x and y from about 22 to 36 times.

The pool of predictors for these forecasts consisted of the variables x, y, z, xx, xy, xz, yy, yz and zz at times t, t - 0.1, t - 0.2, ..., t - 0.9. Using the forward selection method described in the preceding section, Scott found the best three predictors for the first block of 5,000 time steps. These are shown in Table 1 for each variable at time t +.25 in the order they were chosen. It is interesting to note that six of the nine predictors are nonlinear products

Table 1: Predictors for x, y and z in the Lorenz equations chosen by the forward selection method. The first column shows the predicted variable. The next three columns show the predictors in the order selected by the scheme.

Predictand	1	2	3
x (t+.25)	y(t)	yz(t)	yz(t -.1)
y (t+.25)	y(t)	yz(t)	yz(t -.5)
z (t+.25)	z(t -.1)	xx(t -.3)	xy(t -.6)

of the variables. Of these, only the product xy appears in the original governing equations. Moreover, the product xz, which also appears in the governing equations was not chosen by the forward selection method as one of the optimum predictors for the purpose of statistical prediction. Scott used the predictors shown in Table 1 to make 5,000 forecasts on the independent data in each of the other 19 blocks using both linear regression and a neural network. For this purpose, he chose a neural network with one hidden layer consisting of two nodes. The average rms errors and anomaly correlations, plus or minus the standard deviations for both methods, are given in Table 2 for comparison. The predictions of y and z are seen to be poor for both methods, but the prediction of x is much better, particularly using the neural network (capturing 72% of the variance of x). While there is nothing that can be done to improve the linear regression forecast, given the same number of predictors, there are several avenues that can be followed to improve the neural network forecast. For example, we can use different initial weights at each of the existing nodes, increase the number of iterations, increase the number of nodes within the hidden layer and/or increase the number of

Table 2: Rms errors and anomaly correlations for the linear regression and neural network predictions. The first column identifies the predictand. The second and third columns give the rms error and anomaly correlations, respectively, corresponding to the linear regression forecasts. The last two columns give the same measures for the neural network prediction.

	Linear regression		Neural Network	
Predictand	rmse	corr.	rmse	corr.
x (t+.25)	.694±.007	.759±.005	.548±.020	.850±.011
y (t+.25)	1.087±.032	.409±.034	.932±.014	.566±.013
z (t+.25)	.805±.089	.672±.075	.805±.086	.672±.072